DOCUMENT RESUME

ED 319 777 TM 014 995

AUTHOR Witt, Elizabeth A.; And Others
TITLE Recent Trends in Achievement Test Scores: Which Students Are Improving, and on What Levels of Skill Complexity?
PUB DATE Apr 90
NOTE 30p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 17-19, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150) -- Statistical Data (110)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; Comparative Testing; *Educational Trends; Elementary School Students; Elementary Secondary Education; Grade 3; Grade 7; Grade 11; *High Achievement; *Low Achievement; National Programs; *Scores; Secondary School Students; State Programs; Statistical Analysis; Testing Programs; Test Results; Trend Analysis
IDENTIFIERS *Iowa; Iowa Tests of Basic Skills; Iowa Tests of Educational Development; National Assessment of Educational Progress

ABSTRACT

Recent trends in achievement test scores among elementary and secondary school students in Iowa are examined. Form 7 of the Iowa Tests of Basic Skills (ITBS) was used to measure the achievement of students in grades 3 and 7; and Form X-7 of the Iowa Tests of Educational Development (ITED) was used for 11th graders. These grade levels were chosen because they correspond most closely to the age groups for which the disturbing results were received from the National Assessment of Educational Progress (NAEP). ITBS scores of Iowa third and seventh graders in 1978-79 were compared with those of students in the same grades in 1984-85; ITED scores of 11th graders from 1979-80 and 1985-86 were likewise compared. At each grade level in each of the selected school years, data for over 5,000 students were analyzed. Findings of the study are somewhat at odds with those of the NAEP. Specifically, the Iowa study indicated that: (1) although increases in scores are generally greater for low-scoring students, high-scoring students in Iowa showed substantial improvements across the years; (2) no practical differences appeared in the changes in performance for higher-level versus lower-level skills; and (3) students' performance increased on all types of items. Discrepancies with the NAEP study may have resulted from typically higher achievement test scores by Iowa students, the general lack of uniformity of achievement score trends, and differences in data analysis. Twelve data tables are included. (TJH)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED319727

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
ELIZABETH A. WITT
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

# Recent Trends in Achievement Test Scores: Which Students Are Improving, and on What Levels of Skill Complexity?

Elizabeth A. Witt

Myunghee Yun Han

H. D. Hoover

The University of Iowa

Paper Presented at
the Annual Meeting of the
National Council on Measurement in Education
Boston, April, 1990

TM014995

BEST COPY AVAILABLE

## Introduction[1]

In recent years positive trends have been noticed in the achievement test scores of American elementary and secondary school students, indicating a reversal of the general decline which occurred in the 1960s and 1970s. The score declines of those years were evident nationwide at all ability levels, in many subject areas, and on various kinds of tests--most notably on the SAT (Scholastic Aptitude Test) (Breland, 1976) and ACT (American College Testing Program) (Munday, 1976). Similar patterns of decline appeared on the MSAT (Minnesota Scholastic Aptitude Test) (Harnischfeger & Wiley, 1975) and the Iowa tests (Iowa Testing Programs, 1984, 1987); further confirmation of declining achievement was found in the results of the National Assessment of Educational Progress (NAEP) (Harnischfeger & Wiley, 1975). Score declines were most severe among older students (Harnischfeger & Wiley, 1975) and in areas requiring inferential and critical thinking skills (Congressional Budget Office, 1986).

The more recent positive trends have been somewhat less uniform, with scores increasing for various age groups, student ability levels, and subject areas at different points in time. In general, however, the scores of younger children began to increase around 1975, while older groups did not show similar improvements until the early 1980s (The College Board, 1986, 1988; Congressional Budget Office, 1986; Iowa Testing Programs, 1987; National Assessment of Educational Progress [NAEP], 1986), giving the initial impression that gains were occurring only at the lower grade levels.

Recent studies using data from NAEP reports have raised concerns that higher achievement test scores may primarily reflect gains for relatively low-achieving students in low-level skills (Dossey, Mullis, Lindquist, & Chambers, 1988; Forbes, 1985; Mullis & Jenkins, 1988; NAEP, 1986). In other words, students of higher ability have not matched the achievement gains of lower-ability students, and for all students there remains a critical deficiency of higher-order thinking skills (HOTS). These concerns are based primarily on the

---

[1] Material presented in this paper is based on the second author's master's thesis: Han, M. Y. (1988). The Nature of Recent Trends in the Achievement Scores of Elementary and Secondary Students in Iowa. Unpublished master's thesis, The University of Iowa, Iowa City, IA.

tendency (within age group) of lower NAEP achievement classes to show greater improvement than higher achievement classes (Forbes, 1985) and on a lack of significant increase in the percentage of students reaching the highest proficiency levels in mathematics and science (Dossey et al., 1988; Mullis & Jenkins, 1988). Additional concerns about students' higher-order thinking abilities are raised in connection with the more qualitative writing assessment (Applebee, Langer, & Mullis, 1986).

Not all tests, however, have matched the trends apparent in the NAEP data. A comparison of trends on the SAT, the ACT, and other standardized tests at the high school level revealed many discrepancies between tests; score changes over the 1970-1983 interval range from -.26 to .10 standard deviation units (Congressional Budget Office, 1986). Such variation suggests that no single test can provide a definitive portrayal of overall achievement trends.

## Objective

In light of the concerns raised by the NAEP data, the primary objective of this study was to examine recent trends in achievement test scores among elementary and secondary students in Iowa schools. In particular, three questions were addressed:

1. Are the recent improvements in test scores greater for the lower grades than they are for the upper grades?
2. Is there a difference in the magnitude or direction of score changes for students of different ability levels?
3. Is there a difference in the pattern of score changes for items differing in complexity (lower-level versus higher-order thinking skills)?

## Data Sources

### The tests

Form 7 of the Iowa Tests of Basic Skills (ITBS) was used to measure the achievement of students at Grades 3 and 7, and Form X-7 of the Iowa Tests of Educational Development (ITED) was employed for Grade 11. These grade levels were chosen because they correspond

most closely to the age groups for which the NAEP results have been reported.

The annual statewide use of the ITBS and ITED in Iowa produced data particularly useful for examining score changes, as the demographic composition of the test-taking population is extremely stable and the vast majority of Iowa students are assessed. Children at all grade levels are tested over a comprehensive range of subject areas. Moreover, trends in achievement among Iowa students have been similar to those occurring nationally (Congressional Budget Office, 1986).

The Iowa Tests of Basic Skills. The ITBS battery is administered in about 99% of the public and private schools in Iowa; the statewide testing program currently involves between 35,000 and 40,000 students in each of grades 3 through 8. The tests are intended to provide comprehensive and periodic measurement in fundamental academic skills. The concept of fundamental or "basic" skills is not limited to low-level tasks, but covers a variety of abilities--including higher-order cognitive competencies--which are important for a child's success in academic activities and related life skills. The major test areas are:

Vocabulary (V)
Reading Comprehension (R)
Language
- -Spelling (L1)
- -Capitalization (L2)
- -Punctuation (L3)
- -Usage (L4)

Work Study Skills
- -Visual Materials (W1)
- -Reference Materials (W2)

Mathematics
- -Concepts (M1)
- -Problem Solving (M2)
- -Computation (M3)

The ITBS are designed to reflect the continuous nature of skill development. The tests are arranged in levels constructed to reveal progress along a continuous developmental path from kindergarten to Grade 9 (Hieronymus, Lindquist, & Hoover, 1982). The ITBS yield grade equivalent (GE) scores and within-grade percentile ranks for each of three times of year (fall, mid-year, and spring).

The Iowa Tests of Educational Development. In use since 1942, the ITED are designed to measure essential abilities that practically all adolescents and adults must use in daily life, including higher-order competencies. The tests emphasize cognitive skills. They are not closely tied to any curriculum and attempt to measure processes rather than specific content knowledge. Scores reflect intellectual development which occurs as a result of out-of-school experiences as well the effects of school learning (Iowa Testing Programs, 1979). Results are reported in terms of standard scores and percentile ranks.

The ITED assess achievements of high school students (Grades 9-12) in these areas:

English
- Correctness and Appropriateness of Expression (E)
- Spelling (E)

Quantitative - Quantitative Thinking (Q)
Social Studies
- Concepts and Backgrounds (SS)
- Reading Material (SS)

Natural Sciences
- Concepts and Backgrounds (NS)
- Reading Material (NS)

Literature - Interpretation of Literary Material (L)
Vocabulary (V)
Use of Sources of Information (SI)

These tests measure competency in a variety of important skills: recognizing the essentials of good writing, solving quantitative problems, critical analysis of discussions of social issues, understanding nontechnical scientific reports and recognizing sound methods of scientific inquiry, perceiving subtle meanings in literary materials, and making effective use of sources of

information and common tools of learning (Iowa Testing Programs, 1979).

The subjects

ITBS scores of Iowa third and seventh graders in the 1978-79 school year were compared with those of the same grades in 1984-85; ITED scores of eleventh graders from 1979-80 and 1985-86 were likewise compared. Two parallel test forms were administered in a counterbalanced fashion to the same age populations over these six-year periods, with specific forms alternated from year to year. Thus students at each grade level were tested with identical forms at each of the two assessments, permitting direct comparisons of performance.

Only those Iowa schools which administered Form 7 of the ITBS or Form X-7 of the ITED to all regularly enrolled students in Grades 3 and 7 (ITBS) or Grade 11 (ITED) at the same time of year during both assessments were included in the study. Students in these grades are approximately the same age as the children tested by NAEP: 9, 13, and 17. At each grade level in each of the selected school years, data from over 5000 students satisfying the preceding criteria were available. Actual numbers of students are listed by grade level and year in Tables 1-3.

## Methods and Results

Question 1: Are the improvements in scores greater for the lower grades than for the upper grades?

To answer this question, mean scores were computed at each of the three grade levels for each of the two assessments (1978-79/1979-80 and 1984-85/1985-86). Differences between the later and earlier means were calculated. Because the data at the different grade levels were based on tests of differing length, these mean differences were then divided by the standard deviations of the earlier groups. Effect sizes are thus expressed in standard deviation units. In addition, ratios of standard deviations (1984-85 s.d. divided by 1978-79 s.d. for Grades 3 and 7; 1985-86 s.d. divided by 1979-80 s.d. for Grade 11) were computed for the purpose of examining changes in variability.

Significant positive differences ($p < .001$) were found between mean achievement scores for the two assessments in all subject areas at all grade levels. Effect sizes are similar across subject areas and across grade levels. Older students showed as much improvement as younger students. Variability declined slightly (with minor exceptions) in the second assessment. Mean differences and standard deviation ratios are presented by grade level and test area in Tables 1-3; this information is averaged and summarized in Table 4.

Question 2: Is there a difference in score changes for students of different (high versus low) ability levels?

Data from Table 4 showing an increase in means and decrease in variability suggest that low-scoring students have shown more improvement than high-scoring students. This conclusion is substantiated by a comparison (between 1978-79/1979-80 and 1984-85/1985-86 data) of the percentages of students at each of three performance levels: 1) a perfect score, 2) 90% + correct responses, and 3) 30%- correct responses.

In Grades 3 and 7 the proportion of students earning perfect scores increased slightly over the six-year interval. In Grade 11 the percentage of students obtaining perfect scores on the ITED was so small at both test dates that no reasonable comparison could be made. In general, decreases in the percentage of students with a "percent correct" score of 30 or less were greater in magnitude than the corresponding increases in percentage of students with "percent correct" scores of 90 or more. That is, the proportion of students that "moved up" the score scale is greater for the low-scoring group. These data are presented by grade level and test area in Tables 5-7; they are averaged and summarized in Table 8. "Percentage change" figures should be interpreted with caution, since the rates of change depend on differing initial base percentages.

Question 3: Is there a difference in score changes for items of different complexity (lower-level versus higher-level)?

All items on Form 7 of the ITBS and Form X-7 of the ITED were classified as HOTS or non-HOTS items by two independent judges using the skills classifications and guidelines published by the Iowa Testing Programs (1979, 1984, 1988). Determination of the skill complexity required by a test item is by nature a subjective process; the procedures followed in classifying items are described in greater detail in the appendix.

For each subtest the average difficulty levels (p-values) were computed for HOTS and for non-HOTS items. Differences between later (1984-85/1985-86) and earlier (1978-79/1979-80) average p-values were calculated. These values are reported by subject area for each grade level in Tables 9-11. Nearly all differences are positive, evidencing improved performance on both higher-and lower-level skill items for all subtests. (The one exception is for the Reading test at Grade 7, where a slight decline occurred for lower-level items only.) The magnitude of change on HOTS items is similar to that on non-HOTS items. Indeed, aside from those test areas which are composed entirely of lower-level items, the increases are in general slightly greater for HOTS items. Summary data reporting p-values averaged across test areas are given in Table 12.

Summary

A major limitation of this study should be mentioned. There is a possibility that some teachers may have explicitly taught specific test items. The same test forms were administered every other year, and comparisons were made between the first and fourth administrations of the same form. However, it seems unlikely that this actually occurred. The tests were not available to the teachers throughout the year; test materials are returned to the Iowa Testing Programs following each administration. More importantly, school level accountability is not emphasized in the testing of Iowa students. Most of the testing is done in the fall, and test use focuses on diagnosing individual needs for the purpose of improving instruction. Explicitly "teaching to the test" would serve little purpose and would not be expected to occur on a grand

enough scale to markedly affect statewide means. In addition, if instruction were focused on specific items or skills, it seem most likely that this focus would be directed toward the lower-level items, which are more easily taught. Gains in lower-level skills would thus be much greater than those actually found.

Overall, the results of this study are encouraging. In Iowa, during the period of time investigated, achievement test scores improved as much for junior high and high school students as they did for the younger children. The gap between the performance of high-scoring and low-scoring students was somewhat reduced; this reduction appears to have taken place without a decline in the overall performance of high-scoring students and was not at the expense of achievement on higher-order, critical thinking skills.

## Discussion

The past two decades have seen a variety of changes in America's approach to education, including the "back to basics" movement of the 1970s and the more recent return to an emphasis on traditional academic goals. A possible effect of these movements can be seen in the recent upward trends in achievement scores of elementary and secondary school students. This study helps to clarify the nature of those trends. Results are largely congruent with the improvements noted by the NAEP for students of similar ages. However, the findings of this study differ from the NAEP reports (Applebee et al., 1988; Dossey et al., 1988; Mullis & Jenkins, 1988; NAEP, 1986) in two encouraging ways:

1. Although the increase in scores is greater for low-scoring students, high-scoring students in Iowa also showed substantial improvement between 1978-79 (1979-80) and 1984-85 (1985-86) at each grade level studied.
2. This study found no practical differences in the changes in performance for higher- level versus lower-level skills; Iowa students' performance increased on all types of items.

Obviously, given the sample of Iowa students, we cannot be certain of the extent to which the results of this study hold true for students across the nation. Iowa students, on the average, tend to score higher than the national average on standardized achievement tests. This may partially account for the difference between our findings and the NAEP data. If a relatively small percentage of Iowa students are "low achievers," perhaps less school time is required for remediation and concentration on low-level skills, leaving room for greater emphasis on higher-order skills and providing more opportunities for high-scoring students.

Further explanation of the difference from the NAEP findings might be found in the general lack of uniformity of achievement score trends. Even the NAEP assessments show no difference in improvement in lower-order versus higher-order thinking skills for some groups on some tests. In reading, for example, 9-year-olds exhibited similar patterns of score changes at different levels of skill complexity in the 1970s (NAEP, 1981); at all ages over the period 1980-1984, score changes in literal comprehension differed only slightly (not significantly) from those in inferential comprehension (NAEP, 1986).

A difference in this study's approach to data analysis may also be relevant. We have conceptualized higher-order processes in terms of what the item requires of the test taker. The NAEP data are summarized in terms of proficiency level; for most tests, items are not explicitly divided into higher-order versus lower-order classifications. Items of varying complexity may appear at all levels of proficiency. In the NAEP reading tests, although the difficulty of passages increases at higher proficiency levels, attendant item sets include both literal and inferential comprehension questions at all levels (NAEP, 1986). An examination of sample items in math and science indicates that higher proficiency levels are marked by advanced content as well as advanced reasoning skills, and some items reflect simple reasoning or recall (non-HOTS) of advanced material (see for example Dossey et al., 1988, p. 41). High-proficiency items in science, especially, seem to rely on advanced content as well as higher-order processing for their difficulty. Possible improvements in higher-order thinking skills may thus be obscured as HOTS and non-HOTS items are interwoven with proficiency

levels. It would be interesting to see what results the NAEP data would yield if trends in higher-order thinking were examined at the item level.

The NAEP assessments clearly indicate that, as a group, American students have not attained the academic proficiency levels we would like; achievement goals which appear to be reasonable have not yet been met, particularly in math and science. However, it is not clear that this academic shortfall reflects a lack of complex thinking abilities. The results of this study indicate that complex cognitive skills have not fallen by the wayside in Iowa schools. Although Iowa students cannot be said to constitute a nationally representative sample, the pattern of improvements they have shown implies that the renewed emphasis on basic skills and traditional academic goals need not be detrimental to high-achieving students nor require a neglect of higher-order thinking skills.

## References

Applebee, A. N., Langer, J. A., & Mullis, I. V. S. (1986). The writing report card: Writing achievement in American schools (The Nation's Report Card: Report No. 15-W-02). Princeton, NJ: Educational Testing Service.

Applebee, A. N., Langer, J. A., & Mullis, I. V. S. (1988). Who reads best? Factors related to reading achievement in grades 3, 7, and 11 (The Nation's Report Card: Report No. 17-R-01). Princeton, NJ: Educational Testing Service.

Breland, H. M. (1977). The SAT score decline: A summary of related research (Table 1). New York: The College Board.

The College Board. (1986). College bound seniors. Princeton, NJ: The College Board.

The College Board. (1988). College bound seniors: 1988 profile of SAT and achievement test takers. Princeton, NJ: The College Board.

Congressional Budget Office. (1986). Trends in educational achievement. Washington, DC: Congressional Budget Office.

Dossey, J. A., Mullis, I. V. S., Lindquist, M. M., & Chambers, D. L. (1988). The mathematics report card: Are we measuring up? Trends and achievement based on the 1986 national assessment (The Nation's Report Card: Report No. 17-N-01). Princeton, NJ: Educational Testing Service.

Forbes, R. H. (1985). Academic achievement of historically lower-achieving students during the seventies. Phi Delta Kappan, 66, 542-544.

Harnischfeger, A. & Wiley, D. E. (1975). Achievement test score decline: Do we need to worry?. Chicago: Cemrel.

Hieronymus, A. N., Lindquist, E. F., & Hoover, H. D. (1982) Iowa Tests of Basic Skills: Manual for school administrators. Chicago: Riverside.

Iowa Testing Programs. (1979). Iowa Tests of Educational Development: Manual for teachers, counselors, and examiners. Iowa City: University of Iowa.

Iowa Testing Programs. (1984). Mean ITED scores by grade and subtest for the state of Iowa: 1962-present. Unpublished raw data.

Iowa Testing Programs. (1987). Achievement trends in Iowa: 1955-1987. Unpublished manuscript.

Iowa Testing Programs. (1988). Iowa Tests of Basic Skills: Higher order thinking skills supplemental guide. (Limited publication. Available from The Riverside Publishing Company, 8420 Bryn Mawr Avenue, Chicago, IL 60631.)

Mullis, I. V. S., & Jenkins, L. B. (1988). The science report card: Elements of risk and recovery. Trends and achievement based on the 1986 national assessment (The Nation's Report Card: Report No. 17-S-01). Princeton, NJ: Educational Testing Service.

Munday, L. A. (1976, February). Declining admissions test scores (ACT Research Report No. 71). Iowa City: The American College Testing Program.

National Assessment of Educational Progress. (1981). Three national assessments of reading: Changes in performance 1970-1980 (Report No. 11-R-01). Denver: Education Commission of the States.

National Assessment of Educational Progress. (1986). The reading report card: Progress toward excellence in our schools. Trends in reading over four national assessments, 1971-1984 (Report No. 15-R-01). Princeton, NJ: Educational Testing Service.

# APPENDIX

## DEFINITION AND CLASSIFICATION OF SKILL LEVEL

Some test items are easily classified as requiring higher-order thinking skills (HOTS) or as demanding less complex mental processing (non-HOTS). Most educators would agree, for example, that an item which asks a student simply to identify an error of capitalization or punctuation would fall in the non-HOTS category. Many test items, however, are more difficult to categorize. These items must be examined individually and subjectively, and experts may disagree in their classifications. Whether or not they concur on classification, experts share the opinion that some test items may require higher-order thinking functions of some examinees but not of others (Iowa Testing Programs, 1988). For example, one test taker may be able to answer a question correctly by recalling and executing a simple memorized formula while another examinee, who has not learned the formula, must actively reason through the problem. Any HOTS/non-HOTS classification scheme must therefore be regarded as somewhat arbitrary. The following paragraphs describe the criteria by which ITBS and ITED items were categorized as HOTS or non-HOTS for the purposes of this study.

The Iowa Tests of Basic Skills (ITBS). For the ITBS, classification procedures are described in a supplemental guide (Iowa Testing Programs, 1988). ITBS items included in the HOTS category require the exercise of relatively complex cognitive functions. Simple recall of information is insufficient; the student must actively use information in an operative sense to achieve a goal. The general conception of higher-order thinking encompasses the cognitive operations of interpretation, inference, classification, analysis, and comparison.

All items in the ITBS Vocabulary test (V) are classified as non-HOTS.

For the Reading test (R), items based on recognition of literal meaning or comprehension of stated facts or relationships are regarded as non-HOTS. Those that require the student to use inductive reasoning, or to draw conclusions or make predictions on the basis of information given or implied in the passage are designated as HOTS. The content of distractors was also taken into account in the classification process; if two or more response choices can easily be ruled out on the basis of information presented literally in the passage, the item is labeled as

non-HOTS. Presumably examinees may be able to arrive at the correct answer by eliminating literal distractors without exercising complex thinking skills.

For the Language tests (L1-L4), none of the Spelling, Capitalization, or Punctuation items are labeled as HOTS. Although some test takers may engage in complex thinking on some of these items, there is no way of knowing whether a given item was answered correctly by reasoning or by mere recall. Overall the level of thinking demanded by these tests appears to be insufficiently complex, in comparison with the conception of HOTS for other tests, to warrant a HOTS classification. It is assumed that most of the items can be answered by most students via simple recall.

In the current Forms G and H of the ITBS, higher-order thinking skills are required throughout the Usage and Expression (L4) test. Form 7, however, did not contain any "effectiveness of expression" items, which represent the majority of the G and H L4 HOTS items; therefore it was decided for the purposes of this study to label all L4 items as non-HOTS also.

For the Work Study tests (W1-W2), items that require the use of multiple thinking steps in using maps or charts are classified as HOTS. On the other hand, the use of a chart or map to describe or retrieve direct information is designated as non-HOTS.

For the Mathematics tests (M1-M3), items that require multiple comparisons, multiple operations, multiple thought stages, or a complex conversion of information are regarded as HOTS. Even a multi-step item may be classified as non-HOTS, however, if its solution is based on a rote scheme or if it is very similar to questions the student is likely to have encountered during instruction. Such items are considered to require more recall than reasoning.

The item numbers of ITBS Form 7 items classified as HOTS are listed in Table A-1.

The Iowa Tests of Educational Development (ITED). For the ITED, all items on the Vocabulary test (V) are labeled as non-HOTS. Skill classifications for the Social Studies (SS), Natural Science (NS), Literary Materials (L), and Use of Sources of Information (SI) tests are provided in the ITED Manual for Teachers, Counselors, and Examiners (Iowa Testing

Programs, 1979). The classification scheme is structured around three levels: knowledge and comprehension, application of principles and interpretation, and critical analysis. For tests SS, NS, L, and SI, all items in the first level (knowledge and comprehension) are classified as non-HOTS. Items in the remaining two levels are combined and classified as HOTS. These items involve relatively complex tasks such as inferring unstated relationships, predicting outcomes, distinguishing fact from opinion, identifying logical consequences of given propositions, and formulating testable hypotheses.

Items on the ITED tests of Correctness and Appropriateness of Expression (E) and Quantitative Thinking (Q) were examined individually. Those involving mere recognition, mechanical application of rules, or simple, straightforward computation are classified as non-HOTS. Items requiring students to set up multi-step mathematical solutions, draw logical conclusions, judge the appropriateness of a statement's content or style, or make use of contextually implicit information are designated as HOTS.

The item numbers of ITED Form X-7 items classified as HOTS are listed in Table A-2.

TABLE 1

Summary Data 1978-79 and 1984-85

ITBS Form 7, Grade 3

| Test | V | R | L1 | L2 | L3 | L4 | W1 | W2 | M1 | M2 | M3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 84-85 Mean | 22.45 | 30.34 | 22.36 | 20.34 | 17.65 | 17.46 | 24.66 | 24.32 | 19.58 | 14.86 | 25.68 |
| 78-79 Mean | 21.39 | 28.60 | 21.38 | 19.01 | 16.60 | 16.30 | 23.57 | 23.00 | 19.10 | 14.22 | 24.26 |
| Mean Difference | 1.06 | 1.74 | .98 | 1.33 | 1.05 | 1.16 | 1.09 | 1.32 | .48 | .64 | 1.42 |
| *Mean Diff. in σ units* | *.16* | *.19* | *.15* | *.22* | *.18* | *.20* | *.16* | *.17* | *.09* | *.13* | *.18* |
| 84-85 S.D. | 5.70 | 8.21 | 5.81 | 5.49 | 5.76 | 5.43 | 6.09 | 7.20 | 4.80 | 4.90 | 7.51 |
| 78-79 S.D. | 6.53 | 8.97 | 6.57 | 6.11 | 5.92 | 5.61 | 6.66 | 7.61 | 5.34 | 5.01 | 7.95 |
| *S.D. Ratio* | *.87* | *.92* | *.88* | *.90* | *.97* | *.93* | *.91* | *.95* | *.90* | *.98* | *.94* |
| 84-85 N | 5359 | 5360 | 5362 | 5362 | 5359 | 5359 | 5336 | 5336 | 5338 | 5333 | 5334 |
| 78-79 N | 6637 | 6637 | 6435 | 6430 | 6430 | 6431 | 6428 | 6425 | 6636 | 6633 | 6597 |

TABLE 2

Summary Data 1978-79 and 1984-85

ITBS Form 7, Grade 7

| Test | V | R | L1 | L2 | L3 | L4 | W1 | W2 | M1 | M2 | M3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 84-85 Mean | 29.52 | 34.55 | 29.58 | 20.87 | 21.05 | 20.81 | 27.19 | 31.26 | 24.55 | 18.71 | 27.98 |
| 78-79 Mean | 28.35 | 33.47 | 27.87 | 19.89 | 20.03 | 19.76 | 25.74 | 30.31 | 23.97 | 18.03 | 26.29 |
| Mean Difference | 1.17 | 1.08 | 1.71 | .98 | 1.02 | 1.05 | 1.45 | .95 | .58 | .68 | 1.69 |
| *Mean Diff. in Units* | *.14* | *.10* | *.18* | *.16* | *.16* | *.17* | *.16* | *.11* | *.07* | *.11* | *.20* |
| 84-85 S.D. | 7.85 | 9.84 | 8.75 | 5.69 | 5.93 | 5.75 | 9.07 | 7.94 | 7.90 | 6.05 | 8.68 |
| 78-79 S.D. | 8.65 | 10.36 | 9.47 | 6.05 | 6.37 | 6.18 | 8.88 | 8.33 | 7.98 | 6.04 | 8.58 |
| *S.D. Ratio* | *.91* | *.95* | *.92* | *.94* | *.93* | *.93* | *1.02* | *.95* | *.99* | *1.00* | *1.01* |
| 84-85 N | 6795 | 6800 | 6791 | 6790 | 6788 | 6788 | 6783 | 6785 | 6793 | 6792 | 6791 |
| 78-79 N | 7264 | 7264 | 7264 | 7258 | 7256 | 7253 | 7250 | 7252 | 7253 | 7250 | 7219 |

TABLE 3

Summary Data 1979-80 and 1985-86

ITED Form X-7, Grade 11

| Test | E | Q | SS | NS | L | V | SI |
|---|---|---|---|---|---|---|---|
| 85-86 Mean | 44.11 | 19.04 | 36.14 | 32.76 | 27.47 | 19.69 | 27.95 |
| 79-80 Mean | 41.51 | 17.75 | 34.17 | 30.79 | 25.55 | 18.19 | 25.63 |
| Mean Difference | 2.60 | 1.29 | 1.97 | 1.97 | 1.92 | 1.50 | 2.32 |
| *Mean Diff. in σ units* | *.20* | *.16* | *.16* | *.17* | *.21* | *.17* | *.28* |
| 85-86 S.D. | 12.16 | 7.77 | 12.37 | 11.61 | 8.89 | 8.86 | 8.10 |
| 79-80 S.D. | 13.30 | 7.84 | 12.65 | 11.80 | 9.22 | 8.93 | 8.32 |
| *S.D. Ratio* | *.91* | *.99* | *.98* | *.98* | *.96* | *.99* | *.97* |
| 85-86 N | 6321 | 6321 | 6321 | 6321 | 6321 | 6321 | 6321 |
| 79-80 N | 7731 | 7731 | 7731 | 7731 | 7731 | 7731 | 7731 |

TABLE 4

Changes in Average Performance ($\sigma_{units}$)

and Variability (S.D. Ratio) Across Grades

| Grade | 3 | 7 | 11 |
|---|---|---|---|
| Average of Mean Differences in $\sigma_{units}$ | .17 | .14 | .19 |
| Average of S.D. Ratios | .92 | .96 | .97 |

TABLE 5

Percentage of Students Scoring at Extremes

1978-79 and 1984-85, Grade 3

| TEST | | V | R | L1 | L2 | L3 | L4 | W1 | W2 | M1 | M2 | M3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| % of students | 1984-85 | 3.1 | .2 | 3.9 | 3.0 | .7 | .6 | .5 | .3 | .9 | 2.3 | 1.1 |
| getting | 1978-79 | 2.5 | .1 | 3.8 | 2.3 | .5 | .4 | .3 | .3 | 1.2 | 1.8 | .5 |
| perfect scores | Difference | .6 | .1 | .1 | .7 | .2 | .2 | .2 | 0 | -.3 | .5 | .6 |
| % of students | 1984-85 | 27.3 | 10.7 | 27.3 | 25.5 | 11.5 | 11.6 | 11.9 | 12.2 | 14.6 | 13.4 | 12.5 |
| getting | 1978-79 | 25.5 | 8.4 | 25.8 | 20.5 | 8.2 | 8.6 | 10.4 | 9.3 | 15.3 | 10.8 | 9.7 |
| 90% or more | Difference | 1.8 | 2.3 | 1.5 | 5.0 | 3.3 | 3.0 | 1.5 | 2.9 | -.7 | 2.6 | 2.8 |
| items right | *% change* | *7.1* | *27.4* | *5.8* | *24.4* | *40.2* | *34.9* | *14.4* | *31.2* | *-4.6* | *24.1* | *28.9* |
| % of students | 1984-85 | 3.6 | 3.7 | 4.0 | 4.0 | 8.1 | 7.9 | 3.2 | 6.0 | 2.6 | 8.6 | 5.2 |
| getting | 1978-79 | 7.1 | 7.0 | 7.2 | 7.4 | 11.2 | 12.8 | 5.8 | 9.0 | 4.3 | 11.5 | 8.5 |
| 30% or fewer | Difference | -3.5 | -3.3 | -3.2 | -3.4 | -3.1 | -4.9 | -2.6 | -3.0 | -1.7 | -2.9 | -3.3 |
| items right | *% change* | *-49.3* | *-47.1* | *-44.4* | *-45.9* | *-27.7* | *-38.3* | *-44.8* | *-33.3* | *-39.5* | *-25.2* | *-38.8* |

TABLE 6

Percentage of Students Scoring at Extremes

1978-79 and 1984-95, Grade 7

| TEST | | V | R | L1 | L2 | L3 | L4 | W1 | W2 | M1 | M2 | M3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| % of students | 1984-85 | .5 | 0 | 1.3 | .9 | 1.1 | .6 | 0 | .1 | .1 | .6 | .7 |
| getting | 1978-79 | .9 | 0 | .8 | .6 | .9 | .4 | 0 | .1 | 0 | .5 | .4 |
| perfect scores | Difference | -.4 | 0 | .5 | .3 | .2 | .2 | 0 | 0 | .1 | .1 | .3 |
| % of students | 1984-85 | 11.5 | 4.0 | 15.5 | 10.8 | 14.3 | 10.4 | 1.2 | 8.1 | 4.2 | 8.6 | 8.3 |
| getting | 1978-79 | 11.4 | 3.7 | 13.1 | 9.1 | 11.6 | 8.5 | .6 | 6.9 | 3.6 | 6.4 | 5.4 |
| 90% or more | Difference | .1 | .3 | 2.4 | 1.7 | 2.7 | 1.9 | .6 | 1.2 | .6 | 2.2 | 2.9 |
| items right | *% change* | *.9* | *8.1* | *18.3* | *18.7* | *23.3* | *22.4* | *100.0* | *17.4* | *16.7* | *34.4* | *53.7* |
| % of students | 1984-85 | 3.5 | 4.6 | 5.9 | 3.8 | 4.0 | 4.6 | 12.7 | 2.9 | 9.3 | 8.9 | 6.6 |
| getting | 1978-79 | 6.6 | 7.1 | 9.8 | 6.2 | 6.9 | 7.6 | 16.0 | 4.2 | 11.2 | 10.4 | 9.3 |
| 30% or fewer | Difference | -3.1 | -2.5 | -3.9 | -2.4 | -2.9 | -3.0 | -3.3 | -1.3 | -1.9 | -1.5 | -2.7 |
| items right | *% change* | *-47.0* | *-35.2* | *-39.8* | *-38.7* | *-42.0* | *-39.5* | *-20.6* | *-31.0* | *-17.0* | *-14.4* | *-29.0* |

TABLE 7

Percentage of Students Scoring at Extremes

1979-80 and 1985-86, Grade 11

| TEST | | E | Q | SS | NS | L | V | SI |
|---|---|---|---|---|---|---|---|---|
| % of students | 1985-86 | 0 | .4 | 0 | 0 | 0 | .4 | 0 |
| getting | 1979-80 | 0 | .2 | 0 | 0 | 0 | .4 | 0 |
| perfect scores | Difference | 0 | .2 | 0 | 0 | 0 | 0 | 0 |
| % of students | 1985-86 | 4.6 | 6.0 | 6.5 | 3.2 | 3.8 | 5.0 | 3.8 |
| getting | 1979-80 | 3.5 | 5.0 | 5.5 | 2.5 | 2.9 | 4.1 | 2.2 |
| 90% or more | Difference | 1.1 | 1.0 | 1.0 | .7 | .9 | .9 | 1.6 |
| items right | *% change* | *31.4* | *20.0* | *18.2* | *28.0* | *31.0* | *22.0* | *72.7* |
| % of students | 1985-86 | 4.9 | 19.2 | 8.8 | 11.2 | 9.7 | 24.3 | 5.8 |
| getting | 1979-80 | 9.0 | 24.9 | 12.4 | 16.2 | 14.3 | 31.3 | 10.0 |
| 30% or fewer | Difference | -4.1 | -5.7 | -3.6 | -5.0 | -4.6 | -7.0 | -4.2 |
| items right | *% change* | *-45.6* | *-22.9* | *-29.0* | *-30.9* | *-32.2* | *-22.4* | *-42.0* |

TABLE 8

Average Percentage of High-Scoring (90%+ correct)
and Low-Scoring (30%- correct) Students
and Percentage Change in Average Percentage
from Earlier to Later Assessment

| Grade | 3 | 7 | | 11 |
|---|---|---|---|---|
| High-Scoring | | | | |
| 1984-85 | 16.2% | 8.8% | 1985-86 | 4.7% |
| 1978-79 | 13.9% | 7.3% | 1979-80 | 3.7% |
| *% change* | *17.0%* | *20.7%* | *% change* | *28.1%* |
| Low-Scoring | | | | |
| 1984-85 | 5.2% | 6.1% | 1985-86 | 12.0% |
| 1978-79 | 8.4% | 8.7% | 1979-80 | 16.9% |
| *% change* | *-38.1%* | *-29.9%* | *% change* | *-28.9%* |

TABLE 9

1978-79 and 1984-85 P-Values by Skill Level

ITBS Form 7, Grade 3

| TEST | | V | R | L1 | L2 | L3 | L4 | W1 | W2 | M1 | M2 | M3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total | Number of items | 30 | 44 | 30 | 28 | 28 | 27 | 36 | 37 | 28 | 23 | 39 |
| | 1984-85 p-value | .748 | .690 | .745 | .726 | .630 | .647 | .685 | .657 | .699 | .646 | .658 |
| | 1979-78 p-value | .713 | .650 | .713 | .679 | .593 | .604 | .655 | .622 | .682 | .618 | .622 |
| | *Difference* | *.035* | *.040* | *.032* | *.047* | *.037* | *.043* | *.030* | *.035* | *.017* | *.028* | *.036* |
| Higher-Level Skill (HOTS) | Number of items | - | 21 | - | - | - | - | 21 | 9 | 3 | 10 | - |
| | 1984-85 p-value | - | .626 | - | - | - | - | .641 | .680 | .677 | .525 | - |
| | 1978-79 p-value | - | .589 | - | - | - | - | .611 | .626 | .660 | .494 | - |
| | *Difference* | - | *.037* | - | - | - | - | *.030* | *.054* | *.017* | *.031* | - |
| Lower-Level Skill (non-HOTS) | Number of items | - | 23 | - | - | - | - | 15 | 28 | 25 | 13 | - |
| | 1984-85 p-value | - | .748 | - | - | - | - | .747 | .650 | .702 | .739 | - |
| | 1978-79 p-value | - | .706 | - | - | - | - | .716 | .620 | .685 | .714 | - |
| | *Difference* | - | *.042* | - | - | - | - | *.031* | *.030* | *.017* | *.025* | - |

TABLE 10

1978-79 and 1984-85 P-Values by Skill Level

ITBS Form 7, Grade 7

| TEST | | V | R | L1 | L2 | L3 | L4 | W1 | W2 | M1 | M2 | M3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total | Number of items | 43 | 57 | 43 | 31 | 31 | 31 | 52 | 47 | 42 | 30 | 45 |
| | 1984-85 p-value | .687 | .606 | .688 | .673 | .679 | .671 | .523 | .665 | .585 | .624 | .622 |
| | 1978-79 p-value | .659 | .587 | .648 | .642 | .646 | .637 | .495 | .645 | .571 | .601 | .584 |
| | *Difference* | *.028* | *.019* | *.040* | *.031* | *.033* | *.034* | *.028* | *.020* | *.014* | *.023* | *.038* |
| Higher-Level Skill (HOTS) | Number of items | - | 36 | - | - | - | - | 39 | 14 | 7 | 20 | - |
| | 1984-85 p-value | - | .588 | - | - | - | - | .506 | .613 | .510 | .618 | - |
| | 1978-79 p-value | - | .570 | - | - | - | - | .481 | .590 | .496 | .599 | - |
| | *Difference* | - | *.018* | - | - | - | - | *.025* | *.023* | *.014* | *.019* | - |
| Lower-Level Skill (non-HOTS) | Number of items | - | 21 | - | - | - | - | 13 | 33 | 35 | 10 | - |
| | 1984-85 p-value | - | .589 | - | - | - | - | .573 | .687 | .599 | .635 | - |
| | 1978-79 p-value | - | .617 | - | - | - | - | .538 | .668 | .586 | .605 | - |
| | *Difference* | - | *-.028* | - | - | - | - | *.035* | *.019* | *.013* | *.030* | - |

TABLE 11

1979-80 and 1985-86 P-Values by Skill Level

ITED Form X-7, Grade 11

| TEST | | E | Q | SS | NS | L | V | SI |
|---|---|---|---|---|---|---|---|---|
| Total | Number of items | 69 | 36 | 60 | 60 | 46 | 40 | 46 |
| | 1985-86 p-value | .639 | .529 | .601 | .545 | .595 | .492 | .608 |
| | 1979-80 p-value | .602 | .493 | .571 | .514 | .557 | .456 | .558 |
| | *Difference* | *.037* | *.036* | *.030* | *.031* | *.038* | *.036* | *.050* |
| Higher-Level Skill (HOTS) | Number of items | 28 | 13 | 40 | 42 | 41 | - | 14 |
| | 1985-86 p-value | .611 | .512 | .598 | .548 | .593 | - | .553 |
| | 1979-80 p-value | .572 | .481 | .563 | .520 | .553 | - | .486 |
| | *Difference* | *.039* | *.031* | *.035* | *.028* | *.040* | - | *.067* |
| Lower-Level Skill (non-HOTS) | Number of items | 41 | 23 | 20 | 18 | 5 | - | 32 |
| | 1985-86 p-value | .659 | .539 | .609 | .538 | .614 | - | .632 |
| | 1979-80 p-value | .622 | .500 | .588 | .501 | .582 | - | .590 |
| | *Difference* | *.037* | *.039* | *.021* | *.037* | *.032* | - | *.042* |

TABLE 12

Average Differences

Between 1984-85/1985-86 and 1978-79/1979-80 p-Values

for Higher- and Lower-Level Skill Items

| Grade | 3 | 7 | 11 |
|---|---|---|---|
| Higher-Level Skills | .034 | .020 | .040 |
| Lower-Level Skills (test areas containing higher-level skill items) | .029 | .014 | .035 |
| Lower-Level Skills (test areas containing no higher-level skill items) | .038 | .034 | .036 |